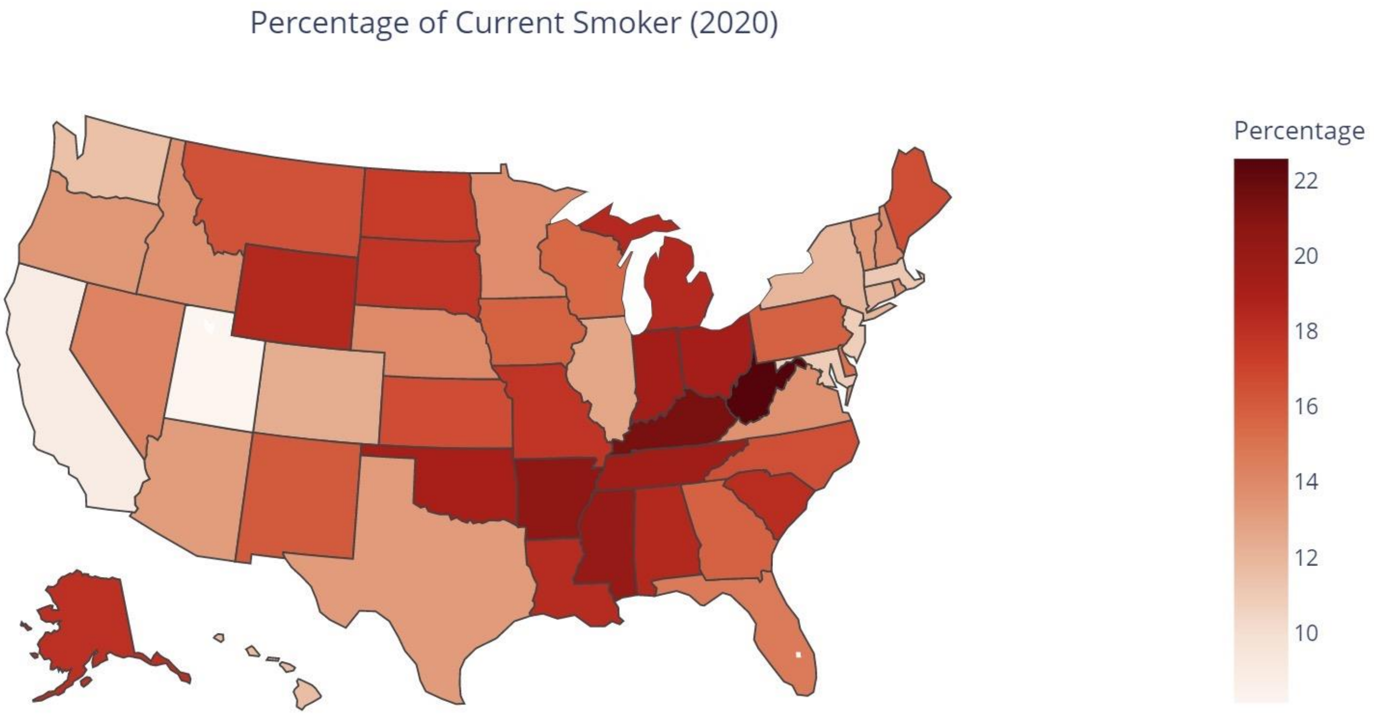
Who's still smoking?

Adult cigarette smoking prevalence in the US and the world



Team 2 report
Chloris Jiang, Jun Xu, Jinsu Zhang, Michael Hu

Preface

Cigarette smoking remains the leading cause of preventable disease, disability, and death in the United States. The high prevalence of cigarette smoking among specific subpopulations, many vulnerable, is one of the most pressing challenges facing the tobacco control community.

Policies restrictions correspond to changing attitudes about smoking and reductions in smoking behavior. Today, 23 states and 493 individual communities in the United States have adopted comprehensive smoke-free laws that prohibit smoking in non-hospitality workplaces, restaurants, and bars.

Here we analyzed several factors that may contributes to the prevalence of tobacco use in the United States and the world. We have broken down our analysis into 3 sections, namely

with one key research question in mind: **Who's still smoking?**

Executive summary

Smoking-related mortality rate in the United States

- Smoking causes about 80% of all deaths from chronic obstructive pulmonary disease (COPD).
- We found that:
    - When holding year constant, mortality rate of COPD is on average 70% lower in female than in male.
    - When ignoring sex, mortality rate of COPD on average increases by 0.14% for every year increase.
    - The odds of COPD mortality rate are 2.8% higher each year in female than for male.
    - About 67% of the variation can be explained by the variation within each county, meaning that within-county differences contributes to 67% of variability of mortality rate

Sociodemographic factors on smoking status in the United States

- Young smokers are decreasing than before.
- Most of smokers are at age 25-65. And differences between groups in age 25-65 are not significant, from Fig 2.4.
- Male smoke more than female.
- People with low income have more probability to be a smoker currently.
- People with higher education have far less prevalence of smoking.
- The differences of level of smoking status, among 2018, 2019, 2020, is not significant. However, that changes fast during 2011 to 2015

Sociodemographic factors on smoking status in the world

- When looking at Female to Male % Smoker ratios, we see a bimodal distribution
    - Ie. Countries cluster into 2 groups
    - One cluster has virtually no female smokers
    - The other cluster has similar female to male smoker %
- These two clusters seem to also be correlated with per capita GDP and the UN's Gender Inequality Index (GII)
    - While not causal, we believe that countries should be careful to not increase female smoking as they develop economically and equally

Smoking-related mortality rate in the United States

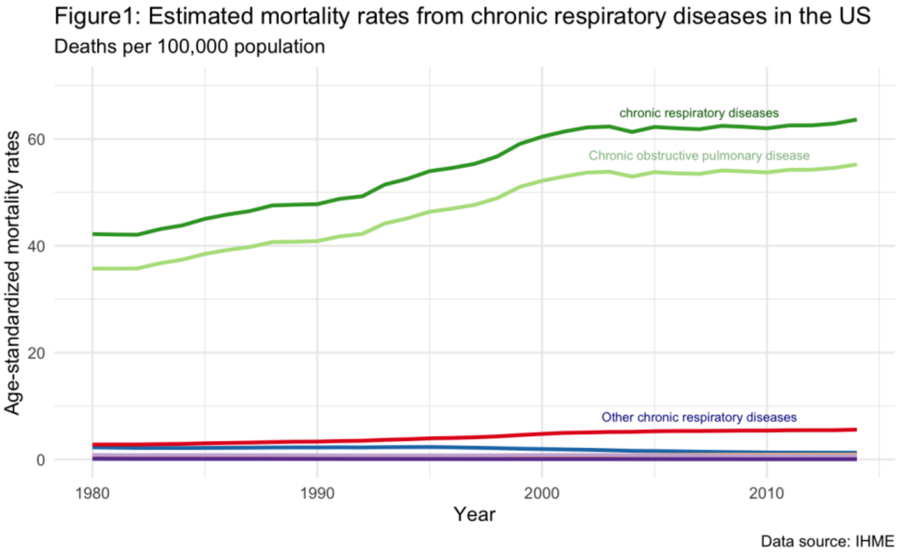


Figure1: Estimated mortality rates from chronic respiratory diseases in the US

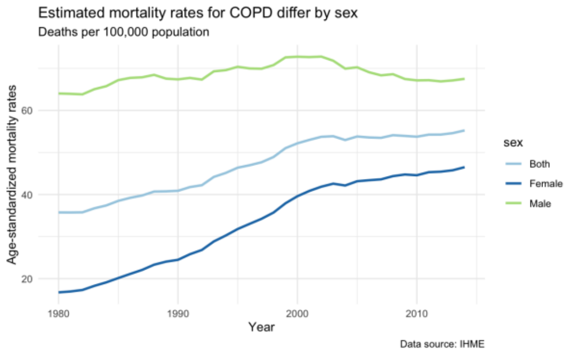
Deaths per 100,000 population

Data source: IHME

Smoking causes about 80% (or 8 out of 10) of all deaths from chronic obstructive pulmonary disease (COPD) Source: CDC . As figure1 shown, it is also one of the deadliest diseases in chronic respiratory diseases.

For the goal of this part of the analysis, we will focus on the factors that can potentially influence mortality rate of chronic obstructive pulmonary disease (COPD). Note that this analysis used data from us_chronic_resp_disease, you can find source data from: IHME

Sex

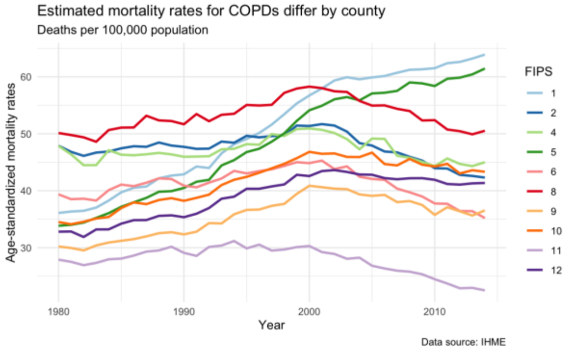

Estimated mortality rates for COPD differ by sex

Deaths per 100,000 population

Data source: IHME

COPD traditionally was considered a man's disease. However, we observed a drastic increase in mortality rate throughout the years for female patients, whereas male mortality rate stayed relatively constant. Research has shown that increasing tobacco consumption among women during the past is linked to the rising prevalence of COPD in women, However, the relationship may be more complex to make causal inference. Link

County


Estimated mortality rates for COPDs differ by county

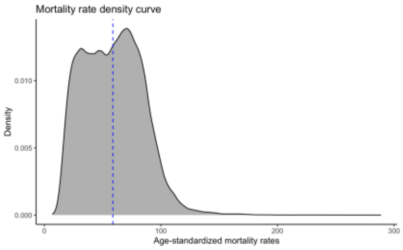
Deaths per 100,000 population

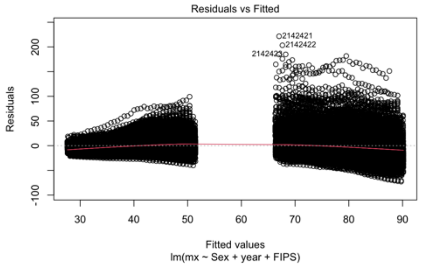
Data source: IHME

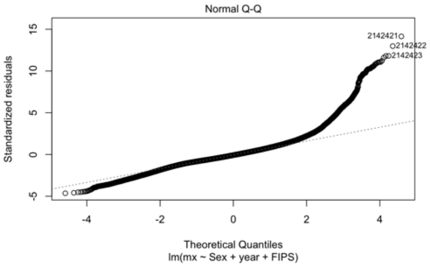
We can see that the mortality rate for COPDs also differs by county. Although we can see that the majorities of counties showed similar trend. You can find more about FIPS and their corresponding area name here. (Note that only a subset of counties are plotted)

1.1.1 Statistical Analysis

We start by checking the assumption of our data. By fitting a simple linear regression and we observe the following diagnostic plots:







- The density plot looks like it might be multimodal and right skewed.
- A heteroscedastic (cone-shaped) pattern on the residual plot, meaning the variance of the residual terms are not constant.
- A heavy-tailed qq plot, meaning our data is right-skewed and isn't normally distributed, in other word, most of the data falls under the larger values.

This is an issue because linear regression assumes that all the observations are independent, which result in uncorrelated and normally distributed residuals.

We know that the data are not independents as they are yearly repeated measurements nested within each of the counties. It is more appropriate to use a mixed effects model rather than a linear regression because having "random factors" in mixed models will help control for the variation coming from each county,

To address these issues, we fit a Log-normal mixed effect model.

Note that some of the key assumptions of Linear Mixed Model are:

- Have a continuous response variable.
- Correctly modelled dependency structure.
- Observations within each subject can be not independent but subjects need to be independent of each other.
- Random effect and with-in unit residual error follow normal distribution.
- Random effect and with-in unit residual error have constant variance.

Log-Normal Mixed Model equation:

$$\log(Y_{ij}) = XB + \mu_i + \epsilon_{ij}$$

$$U_i \sim iid\ N(0, \sigma_U^2)$$

$$\epsilon_{ij} \sim iid\ N(0, \sigma^2)$$

where B represents fixed effect coefficients of year, sex and their interaction, U_i is the random effect intercept for county i, Y_{ij}indicate that each observation j is nested within i.

The reason we do this is because we observed different trends for each county. It would make more sense if county were considered as a random effect in our model. We also transformed the predictor variable to log scale to deal with nonlinearity inherited with the data.

1.1.2 Result:

<u>**Fixed Effect:**</u>

	ESTIMATE	EXP(ESTIMATE)	STD.ERROR	T VALUE	2.5% CI	97.5% CI
(INTERCEPT)	4.3080	74.2918	3.6693e-03	1174.0734	4.3008	4.3152
SEX1	-1.2177	0.2959	1.1721e-03	-1038.9104	-1.2199	-1.2153
YEAR	0.0014	1.0014	4.1913e-05	34.4051	0.0014	0.0015
SEX1:YEAR	0.0284	1.0288	5.9274e-05	479.0027	0.0283	0.0285

Note: The intercept is male mortality rate at year 1980, which gives estimated mortality rate of 74.

All variables are statistically significant according to the confidence intervals (They do not include 0).

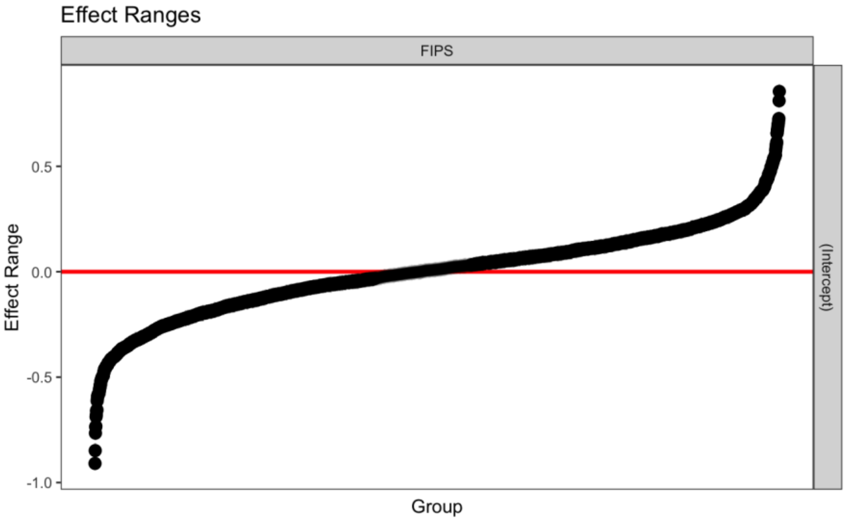
The model for the fixed effect only:

$$Mortality\ rate = 74 + 0.2959 \times sex + 1.0014 \times year + 1.0288 \times sex \times year$$

<u>From the coefficients we can tell that:</u>

- **When holding year constant, mortality rate of COPD is on average 70% lower in female than in male.**
- **When ignoring sex, mortality rate of COPD on average increases by 0.14% for every year increase.**
- **The odds of COPD mortality rate is 2.8% higher each year in female than for male**

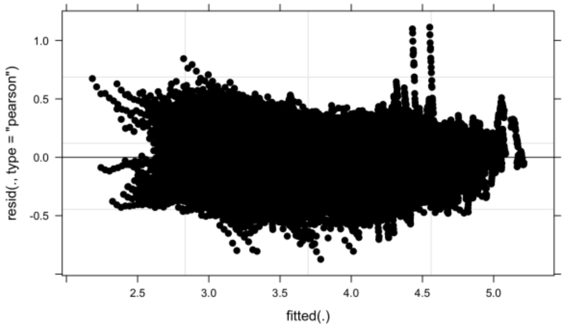
<u>**Random Effect:**</u>

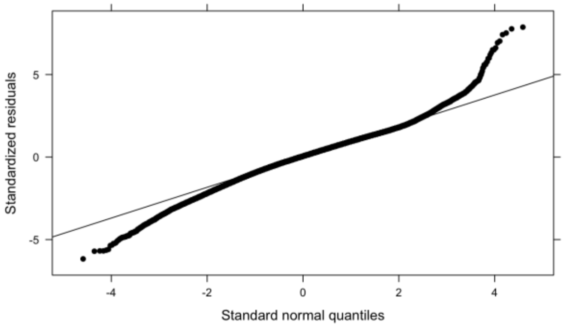


GROUPS	NAME	VARIANCE	STD.DEV.	2.5% CI	97.5% CI
FIPS	(Intercept)	0.04080	0.2020	0.1971	0.2070
RESIDUAL		0.02002	0.1415	0.1411	0.14193

- From the model summary we can calculate the interclass correlation ICC=$\frac{0.04080}{0.02002+0.04080} = 67\%$, which means that about **67% of the variation can be explained by the variation within each county.** This is also confirmed in the dotplot shown above as the confidence intervals for each county didn't overlap. These two tests are to make sure that a mixed model is necessary. However, it is worth noting that there is still 33% of variation not being explained by the random effect, suggesting model have areas of improvement.

1.1.3 Model Diagnostic





The residual plot looks approximately random which indicates homoscedasticity (constant variance), the qq plot however, showed slight heavy tail distribution suggest that there may be other distribution that can fit the data better.

1.1.4 Limitation:

- This model can only be used to interpret the effect of variables on the mortality rate of a certain diseases, which is not entirely smoking related. More analysis on more data needs to be done to conclude smoking-related mortality.
- Mixed model is a powerful model that have interpretability; however, it is easy to build plausible models that are too complex for the data to support and create issue around model reliability.
- More distributional assumptions need to be made on mixed model. Here I assumed log-normal distribution, however based on the diagnostic plot, there may be other model that can fit the data better.

Sociodemographic factors analysis on smoking status in the United States

The dataset, tobacco_use_us, is extracted from BRFSS surveys which collect information about tobacco using in US, from 2011 to 2020. There include three questions:

1. Are you a current smoker? [Yes, No]
2. Do you currently use chewing tobacco, snuff, or snus? [Every day, Not at all, Some days]
3. Four level smoking status: [Never smoked, Smoke everyday, Former smoker, Smoke some days]

Since all the data was extracted from the original system, the measurement of each answer is the proportion of each group. However, we find there exist too much Nan in the group of ethnicities which need to be abandoned. Others are shown in the table.

GENDER	FEMALE	MALE				
AGE	18-24	25-34	35-44	45-54	55-64	65+
INCOME	<15000	15000-24999	25000-34999	35000-49999	50000+	
EDUCATION	<H.S.	HS or GED	post-HS	College graduate		

Moreover, we can only use one factor models for those groups, because different groups have no interaction.

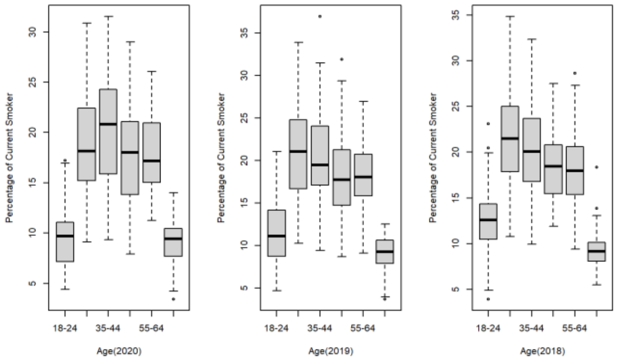
Besides, we are looking for features in those plots, such as Fig 2.1, thus the equality of variance in the groups. Moreover, we are looking for evidence of skewness showing a lack of normality, which might suggest a transformation of the response. Finally, we need to check the significance of difference between each group. The method will be elaborated in section 1.5.

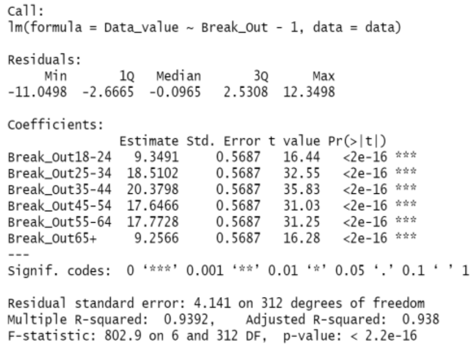
Moreover, we can only use one factor models for those groups, because different groups have no interaction.

Besides, we are looking for features in those plots, such as Fig 2.1, thus the equality of variance in the groups. Moreover, we are looking for evidence of skewness showing a lack of normality, which might suggest a transformation of the response. Finally, we need to check the significance of difference between each group. The method will be elaborated in section 1.5.

2.1.1 Age vs. Percentage (Current Smoker)[1]

Here we want to figure out patterns in percentage of current smoker with respect to 5 age groups. As you can see in Fig 2.1,





```
Call:
lm(formula = Data_value ~ Break_Out - 1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.0498  -2.6665  -0.0965   2.5308  12.3498 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
Break_Out18-24   9.3491     0.5687   16.44   <2e-16 ***
Break_Out25-34  18.5102     0.5687   32.55   <2e-16 ***
Break_Out35-44  20.3798     0.5687   35.83   <2e-16 ***
Break_Out45-54  17.6466     0.5687   31.03   <2e-16 ***
Break_Out55-64  17.7728     0.5687   31.25   <2e-16 ***
Break_Out65+     9.2566     0.5687   16.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.141 on 312 degrees of freedom
Multiple R-squared:  0.9392,    Adjusted R-squared:  0.938 
F-statistic: 802.9 on 6 and 312 DF,  p-value: < 2.2e-16
```

[1] Note: The conclusions of each section are bolded.

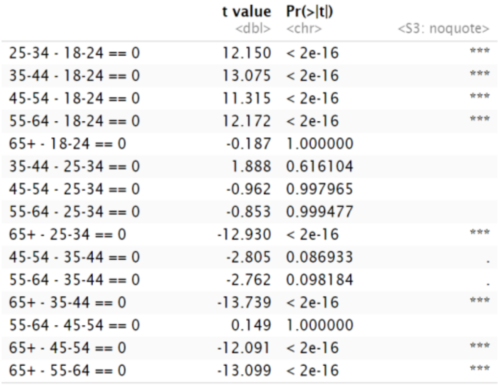
Fig 2.1 Boxplot on the percentage of current smoker on different age group, in 2020, 2019, 2018.

Fig 2.2 Example of the summary of one factor model with respect to percentage of current smoker on age group, in 2020

The results of one factor model indicate the estimated percentage of current smoker in each group, listed in Table 2.3.

	2020	2019	2018
18-24	9.35% ***	11.61% ***	12.72% ***
25-34	18.51% ***	20.86% ***	21.52% ***
35-44	20.38% ***	20.66% ***	20.23% ***
45-54	17.65% ***	18.29% ***	18.70% ***
55-64	17.77% ***	18.24% ***	18.28% ***
65+	9.26	9.23% **	9.42% ***

Table 2.3 Estimated percentage of current smoker in each age group.



```
                    t value  Pr(>|t|)
                      <dbl>  <chr>             <S3: noquote>
25-34 - 18-24 == 0   12.150  < 2e-16                     ***
35-44 - 18-24 == 0   13.075  < 2e-16                     ***
45-54 - 18-24 == 0   11.315  < 2e-16                     ***
55-64 - 18-24 == 0   12.172  < 2e-16                     ***
65+ - 18-24 == 0     -0.187  1.000000
35-44 - 25-34 == 0    1.888  0.616104
45-54 - 25-34 == 0   -0.962  0.997965
55-64 - 25-34 == 0   -0.853  0.999477
65+ - 25-34 == 0    -12.930  < 2e-16                     ***
45-54 - 35-44 == 0   -2.805  0.086933                      .
55-64 - 35-44 == 0   -2.762  0.098184                      .
65+ - 35-44 == 0    -13.739  < 2e-16                     ***
55-64 - 45-54 == 0    0.149  1.000000
65+ - 45-54 == 0    -12.091  < 2e-16                     ***
65+ - 55-64 == 0    -13.099  < 2e-16                     ***
```

Fig 2.4 Pairwise comparison by Tamhane test, in 2020.

From Table 2.3 and Fig 2.4, we can conclude that:

- **Young smokers are becoming less than before.**
- **Most of smokers are at age 25-65. And differences between groups in age 25-65 are not significant, from Fig 2.4.**

2.1.2 Gender vs. Percentage (Current Smoker)

Intuition tells us that men will have more proportion to smoke; however, it needs to be checked.

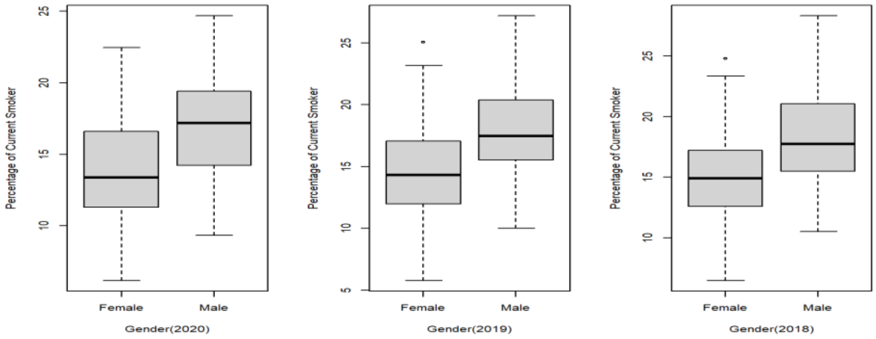


Fig 2.5 Boxplot on the percentage of current smoker with respect to different gender, in 2020, 2019, 2018.

	2020	2019	2018
Female	13.76% ***	14.66% ***	14.84% ***
Male	16.96% ***	17.93% ***	18.36% ***

Table 2.6 Estimated percentage of current smoker with respect to different gender, in 2020, 2019, 2018.

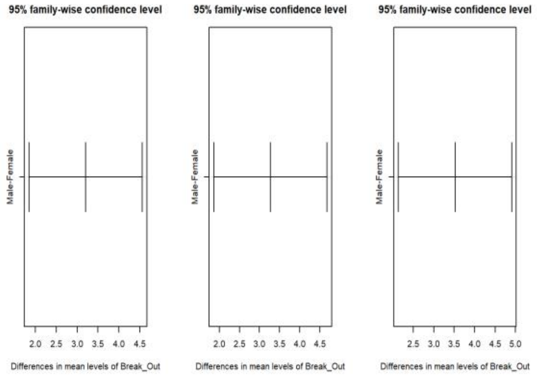


Fig 2.7 Pairwise comparison between different gender, in 2020,2019,2018, by Tukey's HSD test.

From Table 2.6 and Fig 2.5, 2.7, we can conclude that: **Male smoke more than female.**

2.1.3 Income vs. Percentage (Current Smoker)

We wonder if there exist patterns with respect to income, and here illustrates something interesting that we find.

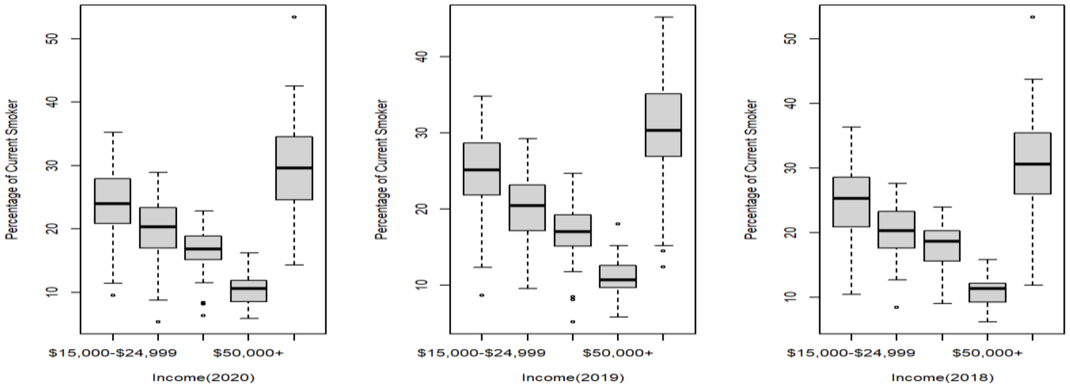
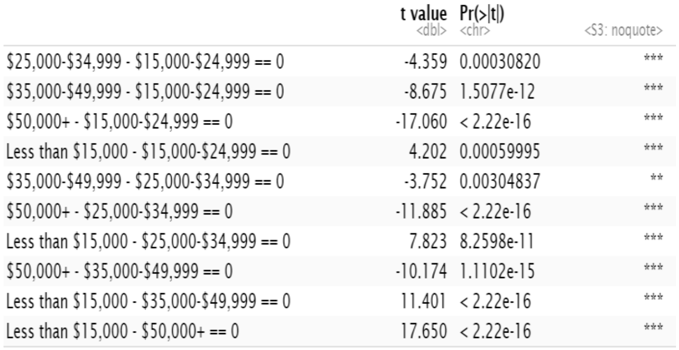


Fig 2.8 Boxplot on the percentage of current smoker with respect to different income groups, in 2020, 2019, 2018. (The right bar is of 'Less than $15,000')

	2020	2019	2018
Less than $15,000	29.49% ***	30.57% ***	30.86% ***
$15,000-$24,999	24.16% ***	24.85% ***	24.81% ***
$25,000-$34,999	19.69% ***	20.11% ***	20.43% ***
$35,000-$49,999	16.44% ***	17.15% ***	18.06% ***
$50,000+	10.44% ***	11.05% ***	10.98% ***



```
                                              t value  Pr(>|t|)
                                                <dbl>  <chr>          <S3: noquote>
$25,000-$34,999 - $15,000-$24,999 == 0         -4.359  0.00030820     ***
$35,000-$49,999 - $15,000-$24,999 == 0         -8.675  1.5077e-12     ***
$50,000+ - $15,000-$24,999 == 0               -17.060  < 2.22e-16     ***
Less than $15,000 - $15,000-$24,999 == 0        4.202  0.00059995     ***
$35,000-$49,999 - $25,000-$34,999 == 0         -3.752  0.00304837     **
$50,000+ - $25,000-$34,999 == 0               -11.885  < 2.22e-16     ***
Less than $15,000 - $25,000-$34,999 == 0        7.823  8.2598e-11     ***
$50,000+ - $35,000-$49,999 == 0               -10.174  1.1102e-15     ***
Less than $15,000 - $35,000-$49,999 == 0       11.401  < 2.22e-16     ***
Less than $15,000 - $50,000+ == 0              17.650  < 2.22e-16     ***
```

Table 2.9 Estimated percentage of current smoker with respect to different income groups, in 2020, 2019, 2018.

Fig 2.10 Pairwise comparison by Tamhane test, in 2020.

From Table 2.8 and Fig 2.9, 2.10, we can conclude that: **people with low income have more probability to be a smoker currently.**

2.1.4 Education vs. Percentage (Current Smoker)

In this section we also find some unpredictable pattern for percentage of current smoker with respect to education attained.

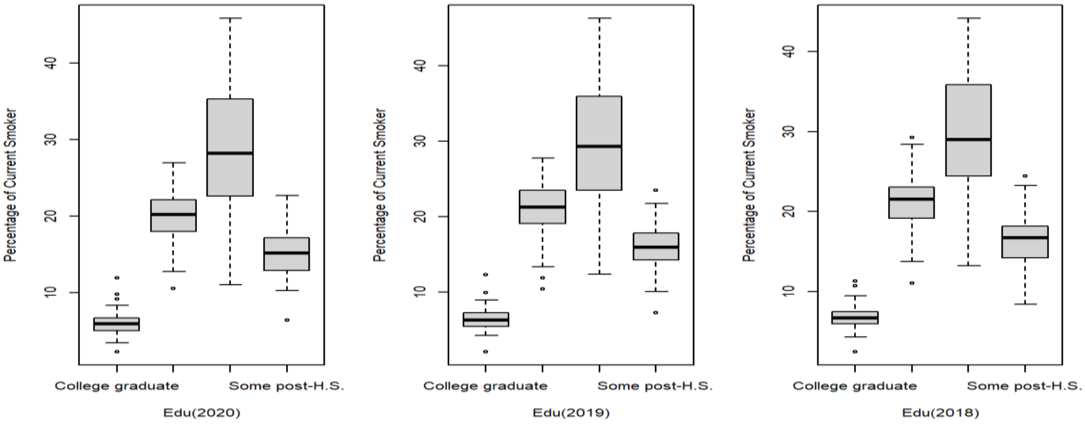
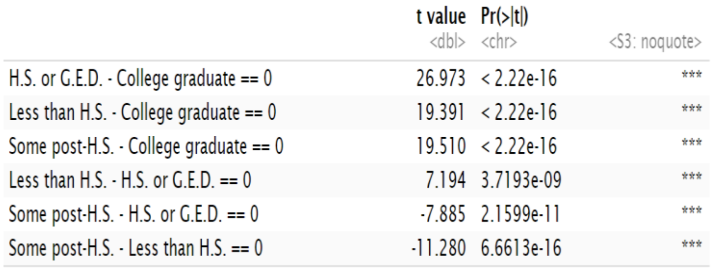


Fig 2.11 Boxplot on the percentage of current smoker with respect to different education attained groups, in 2020, 2019, 2018. This figure is somewhat confusing. The middle two bars are ‘H.S. or G.E.D.’ and ‘Less than H.S.’.

	2020	2019	2018
College graduate	5.81% ***	6.34% ***	6.59% ***
H.S. or G.E.D	19.69% ***	20.59% ***	20.81% ***
Some post-H.S.	14.73% ***	15.89% ***	16.11% ***
Less than H.S.	27.39% ***	27.89% ***	28.40% ***

Table 2.12 Estimated percentage of current smoker with respect to different education attained groups, in 2020, 2019, 2018.



	t value <dbl>	Pr(>|t|) <chr>	<S3: noquote>
H.S. or G.E.D. - College graduate == 0	26.973	< 2.22e-16	***
Less than H.S. - College graduate == 0	19.391	< 2.22e-16	***
Some post-H.S. - College graduate == 0	19.510	< 2.22e-16	***
Less than H.S. - H.S. or G.E.D. == 0	7.194	3.7193e-09	***
Some post-H.S. - H.S. or G.E.D. == 0	-7.885	2.1599e-11	***
Some post-H.S. - Less than H.S. == 0	-11.280	6.6613e-16	***

Fig 2.13 Pairwise comparison by Tamhane test, in 2020.

From Table 2.11 and Fig 2.12, 2.13, we can conclude that:

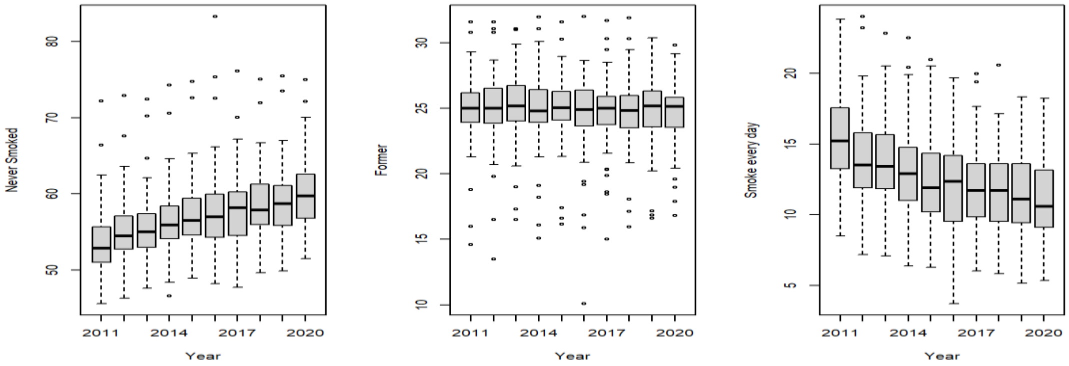
- **people with higher education have far less prevalence of smoking.**

2.1.5 Conclusion

Here we summarize our conclusions of this section:

- **Young smokers are decreasing than before.**
- **Most of smokers are at age 25-65. And differences between groups in age 25-65 are not significant, from Fig 2.4.**
- **Male smoke more than female.**
- **People with low income have more probability to be a smoker currently.**
- **People with higher education have far less prevalence of smoking.**

2.2 Changes in Four Level Smoking Status Over Years



Here we plot the percentage of never smoked people, former smoker, daily smoker, and compare them among years from 2011 to 2020.From QQ-plot of these three models, there exists non-constant variance. Thus, we use Wilcoxon test to examine the differences between specific two years. And the results indicate that there is no significant differences among 2017, 2018, 2019, and 2020. But there exists significant differences for percentage of never smoked people between 2017 and 2020, which justifies the percentage of people who never smoke is increasing. Moreover, you may notice that the level of people smoking decreases fast during 2011 to 2015.

3 Appendix Method

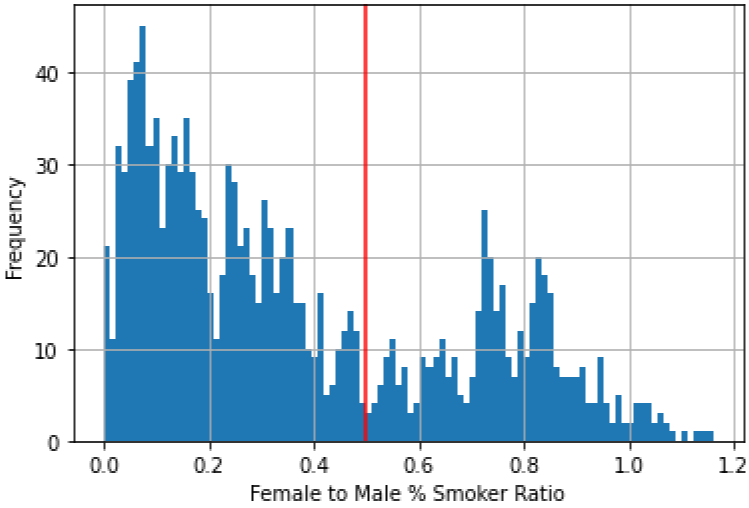
We are looking for features in the plots, such as Fig 2.1, thus the equality of variance in the groups. Now, fit the one factor model in this section, and each factor has several levels. After then, we apply analysis of variance to find if there exists significant difference between the factors, while the test does not tell us which level are different from others.

Before we go on, the assumption of normal distribution in one factor model needs to be checked. Therefore, we plot the residuals and fitted values and make the Q-Q plot of the residuals. If there exists heteroscedasticity, we use box cox method to transform the response. And log(y/(1-y)) is always used for proportion response, which can also be checked by box cox method.

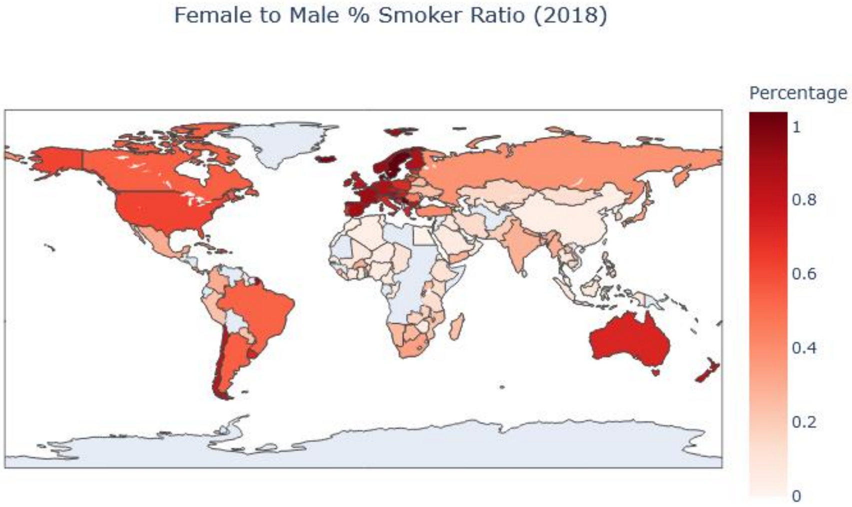
After checking the assumption of normal distribution, we get the final model. After then, the assumption of homogeneity of the error variance need to be examined using Levene's test. While detecting some difference in the levels of factor, the concentration will be which levels is significant from another. For the situation of homogeneity, we use Tukey's honest significant difference (HSD) to calculate 95% confidence intervals for the pairwise differences. Otherwise, for heterogeneity, we use Tamhane's T2 comparison test for data with unequal variances.

Sociodemographic factors analysis on smoking status in the world

3.1.1 Bimodal Distribution of Female to Male Smokers



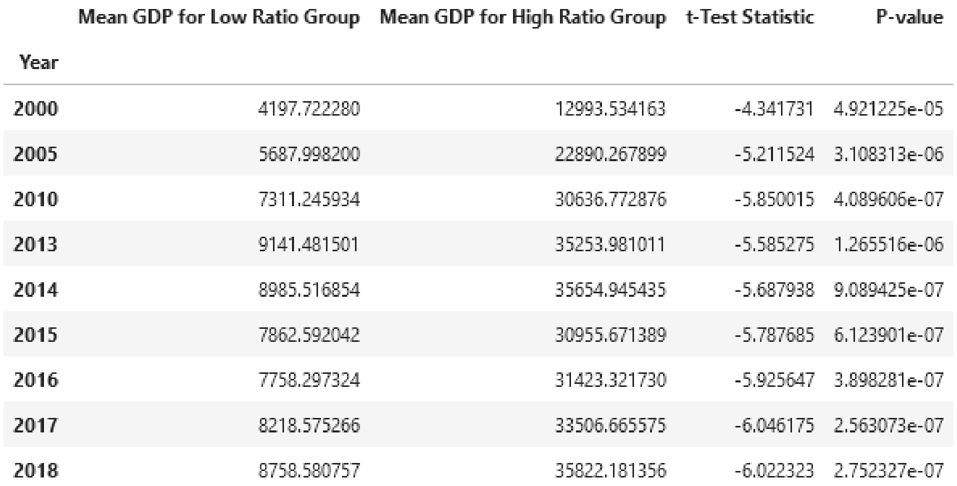
While looking for data anomalies of countries that had high rates of female smokers, we found that there was a large cluster of countries where many women smoked even compared to men. We will define the cluster of countries that has a Female to Male Smoker % Ratio > .5 as the “high ratio group” and <.5 as the “low ratio group”. We decided to investigate further to see if there was any geographic connection as shown below.



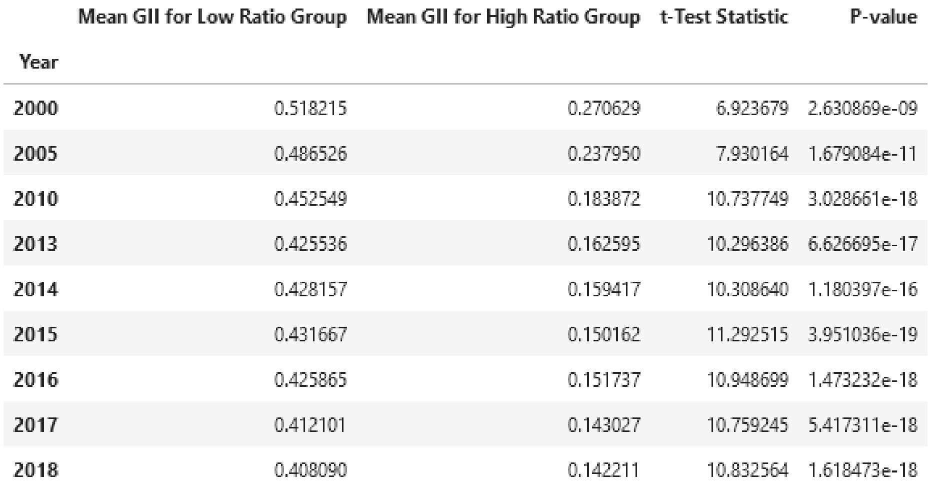
Grouping by continent as well it saw that most (>50%) of high ratio group countries came from Europe and likewise (>50%) of low ratio group countries were from Africa and the Western Pacific. This led us to think that there may be economic factors or women’s rights involved as Europe is highly developed.

3.1.2 Comparison of GDP and GII between low and high ratio groups

We ran a two-way t-Test with non-constant variance (since the two groups did not have the same variance) and saw that for every year, both the GDP and GII were significantly different between the groups. Specifically, countries with a high proportion of female smokers had significantly higher GDP and lower GII, and vice versa.

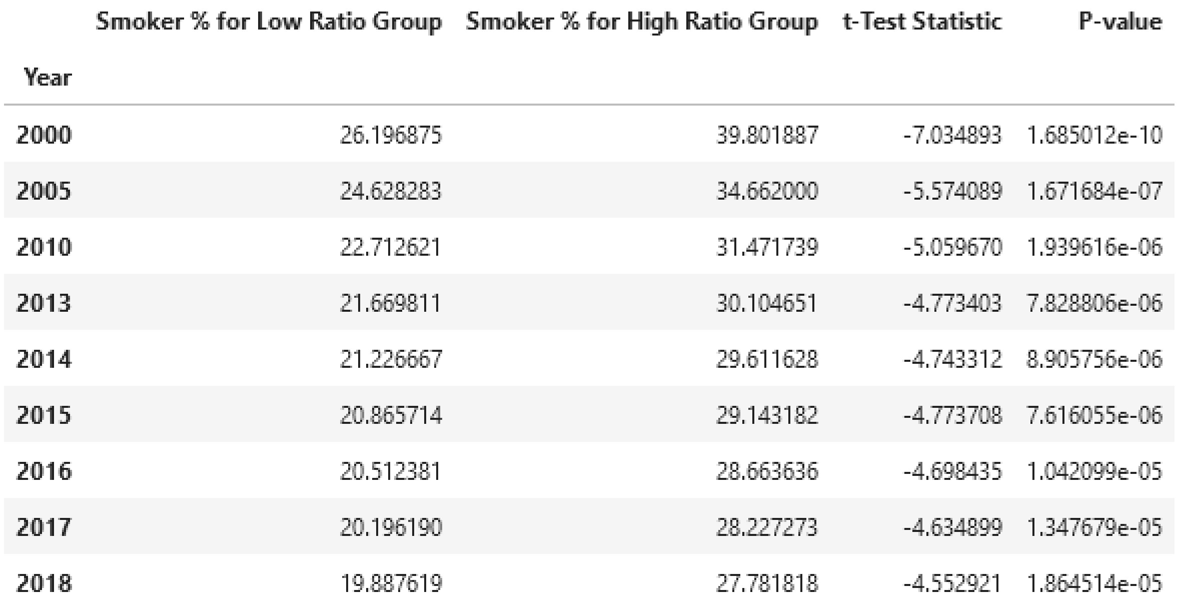


	Mean GDP for Low Ratio Group	Mean GDP for High Ratio Group	t-Test Statistic	P-value
Year				
2000	4197.722280	12993.534163	-4.341731	4.921225e-05
2005	5687.998200	22890.267899	-5.211524	3.108313e-06
2010	7311.245934	30636.772876	-5.850015	4.089606e-07
2013	9141.481501	35253.981011	-5.585275	1.265516e-06
2014	8985.516854	35654.945435	-5.687938	9.089425e-07
2015	7862.592042	30955.671389	-5.787685	6.123901e-07
2016	7758.297324	31423.321730	-5.925647	3.898281e-07
2017	8218.575266	33506.665575	-6.046175	2.563073e-07
2018	8758.580757	35822.181356	-6.022323	2.752327e-07



	Mean GII for Low Ratio Group	Mean GII for High Ratio Group	t-Test Statistic	P-value
Year				
2000	0.518215	0.270629	6.923679	2.630869e-09
2005	0.486526	0.237950	7.930164	1.679084e-11
2010	0.452549	0.183872	10.737749	3.028661e-18
2013	0.425536	0.162595	10.296386	6.626695e-17
2014	0.428157	0.159417	10.308640	1.180397e-16
2015	0.431667	0.150162	11.292515	3.951036e-19
2016	0.425865	0.151737	10.948699	1.473232e-18
2017	0.412101	0.143027	10.759245	5.417311e-18
2018	0.408090	0.142211	10.832564	1.618473e-18

What was shocking to us was that when countries with high proportions of female smokers seemingly just had more smokers. As shown below a two-way t-test (also with non-constant variance) shows that every year, countries with a high ratio of female smokers just had significantly higher smoker % than lower ratio countries.



	Smoker % for Low Ratio Group	Smoker % for High Ratio Group	t-Test Statistic	P-value
Year				
2000	26.196875	39.801887	-7.034893	1.685012e-10
2005	24.628283	34.662000	-5.574089	1.671684e-07
2010	22.712621	31.471739	-5.059670	1.939616e-06
2013	21.669811	30.104651	-4.773403	7.828806e-06
2014	21.226667	29.611628	-4.743312	8.905756e-06
2015	20.865714	29.143182	-4.773708	7.616055e-06
2016	20.512381	28.663636	-4.698435	1.042099e-05
2017	20.196190	28.227273	-4.634899	1.347679e-05
2018	19.887619	27.781818	-4.552921	1.864514e-05

Conclusion

From our analysis, we observed prevalence of cigarette smoking varies by specific subpopulations. Some subpopulations may be vulnerable to tobacco product than others. One of the subpopulations that are increasingly susceptible to tobacco is female. We conclude that subpopulation-based understanding of demographic differences and disparities in smoking is critical to improvement of research design, intervention objectives, and public health policy on smoking in those subpopulations.

Some of the policy-informing ideas we had are:

- Creating and advertising more subpopulation-specific anti-smoking advertisements.
- Penalize smokers with high health insurance premium.
- Providing program-specific training for teachers in schools to educate young adults about harm in smoking.

References

Wikimedia Foundation. (2021, November 17). *List of smoking bans in the United States*. Wikipedia. Retrieved November 21, 2021, from https://en.wikipedia.org/wiki/List_of_smoking_bans_in_the_United_States.

Centers for Disease Control and Prevention. (2021, October 29). *Health effects of cigarette smoking*. Centers for Disease Control and Prevention. Retrieved November 21, 2021, from https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm.

United Nations Development Programme. Gender Inequality Index (GII) Retrieved from http://hdr.undp.org/en/content/gender-inequality-index-gii

The World Bank. GDP per capita (current US$) Retrieved from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD